Supplementary Material for "Data-directed RNA secondary structure prediction using probabilistic modeling"

Fei Deng, Mirko Ledda, Sana Vaziri and Sharon Aviran

Department of Biomedical Engineering and Genome Center,

University of California at Davis, Davis, California, 95616, USA.

Supplementary Methods

Implementation of RNAprob-3

Before giving implementation details, it is worth noting that each helix has two ends - one internal (*type-1 helix-end*) and the other external (*type-2 helix-end*) (Fig. S1). It is straightforward to verify that i-j is a type-1 pair when it closes a hairpin/bulge/internal/multi-branch loop.

In RNAstructure [1, 2], an $N \times N$ array V is used, where $V(i, j)$ is used to represent the MFE of all admissible structures in the subsequence from i to j (denoted by S_{ij}), given that i pairs with

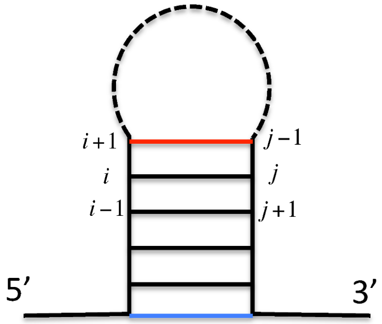

Figure S1: Schematic illustration of type-1 (red line) and type-2 helix-end (blue line).

j. Specifically, $V(i,j)$ is the minimum of the following four terms:

$$V(i,j) = min\{VH(i,j), VI(i,j), VM(i,j), VS(i,j)\},$$

where $VH(i,j)$ is the NNTM free energy of a hairpin loop closed by i-j, $VI(i,j)$ (respectively, $VM(i,j)$) is the MFE over all admissible structures for S_{ij} on the premise that i-j closes a bulge/internal (respectively, multi-branch) loop, and $VS(i,j)$ is the MFE over all admissible structures for S_{ij} satisfying the condition that i-j is stacked on $(i+1)$-$(j-1)$. In addition, two other $N \times N$ arrays, W and WM, and three $1 \times N$ arrays, i.e. $W3$, $W5$ and Wca, are used in the recursion of the dynamic programming algorithm [2].

To implement RNAprob-3, we introduce a new array V_1 of size $N \times N$, such that

$$V_1(i,j) = min\{VH(i,j), VI(i,j), VM(i,j)\} + \Delta G'_i|_{helix-end} + \Delta G'_j|_{helix-end} + \sum_{i<k<j} \Delta G'_k|_{unpaired}.$$

Additionally, we introduce an array V_2 of size $N \times N$, such that

$$V_2(i,j) = VS(i,j) + \Delta G'_i|_{helix-end} + \Delta G'_j|_{helix-end}.$$

Intuitively, $V_1(i,j)$ accounts for all cases where i-j is a type-1 helix-end. In such cases, we are certain about the structural context of i-j. On the other hand, $V_2(i,j)$ accounts for type-2 helix-ends, whose context might change during future sequence extension; in other words, they may become stacked within folds of longer sequences. With V_1 and V_2, we can simply set $V(i,j) = min\{V_1(i,j), V_2(i,j)\}$, which reflects a temporary assumption that i-j is helix-end. Later on, we explicitly check if it is helix-end. The computations of $W(i,j)$, $WM(i,j)$, $W3(i)$, $W5(i)$ and $Wca(i)$ follow from [2], except that $\Delta G'_k|_{unpaired}$ is added for each unpaired base k $(i < k < j)$. We next detail how to calculate $V_2(i,j)$.

In RNAstructure, $VS(i,j)$ is recursively computed as follows:

$$VS(i,j) = V(i+1, j-1) + \Delta G_{stack}(closed\ by\ i\text{-}j\ and\ (i+1)\text{-}(j-1)).$$

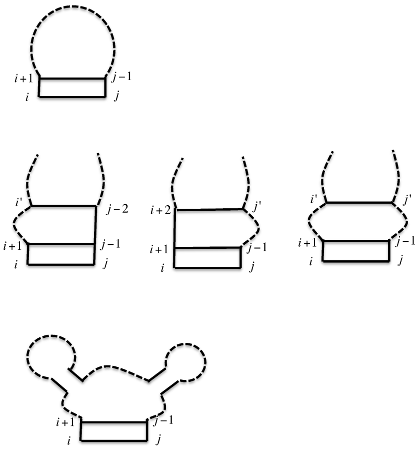

Figure S2: Illustration of Case 1.

Therefore, for RNAprob-3, we have the following general recursion relationship:

$$V_2(i,j) = V(i+1,j-1) + \Delta G_{stack}(\textit{closed by i-j and } (i+1)\text{-}(j-1)) \; + \Delta G'_i|_{helix-end} + \Delta G'_j|_{helix-end}.$$

Note that pseudo-energy terms for $(i+1)$ and $(j-1)$ are no longer included (in contrast to RNAlin's implementation), as they were accounted for under $V(i+1, j-1)$. At this point, when i-j forms, we look back to $(i+1)$-$(j-1)$, to check whether it is helix end or stacked. In essence, we look back to the states of $i+2$ and $j-2$. If $(i+1)$-$(j-1)$ is stacked, then we need to adjust the pseudo-energy terms for $i+1$ and $j-1$ (Fig. S3). More formally, to compute $V_2(i,j)$, we distinguish between two cases:

Case 1: $(i+1)$-$(j-1)$ is a type-1 helix-end, which implies that it closes a hairpin/bulge/internal/ multi-branch loop (Fig. S2). In this case, $V(i+1, j-1) = V_1(i+1, j-1)$. Let $A = V_1(i+1, j-1) + \Delta G_{stack}(\textit{closed by i-j and } (i+1)\text{-}(j-1))$. To this end, we have $V_2(i,j) = A + \Delta G'_i|_{helix-end} + \Delta G'_j|_{helix-end}$.

Case 2: $(i+1)$-$(j-1)$ is stacked onto $(i+2)$-$(j-2)$ (Fig. S3). In this case, $V(i+1, j-1)$ can be computed as $V_2(i+1, j-1) + \Delta(i+1, j-1)$, where $\Delta(i+1, j-1) = \Delta G'_{i+1}|_{stacked} + \Delta G'_{j-1}|_{stacked} - \Delta G'_{i+1}|_{helix-end} - \Delta G'_{j-1}|_{helix-end}$. The term $\Delta(i+1, j-1)$ makes sure that the correct pseudo-energy terms are assigned to $i+1$ and $j-1$, given the fact that $\Delta G'_{i+1}|_{helix-end} + \Delta G'_{j-1}|_{helix-end}$ is applied during the calculation of $V_2(i+1, j-1)$, while $\Delta G'_{i+1}|_{stacked} + \Delta G'_{j-1}|_{stacked}$ is expected.

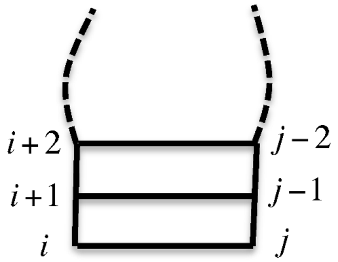

Figure S3: Illustration of Case 2.

Similar to Case 1, let $B = V_2(i+1, j-1) + \Delta(i+1, j-1) + \Delta G_{stack}(\textit{closed by i-j and } (i+1)\text{-}(j-1))$, we have $V_2(i, j) = B + \Delta G'_i|_{helix-end} + \Delta G'_j|_{helix-end}$. Consequently, we have $V_2(i, j) = min\{A, B\} + \Delta G'_i|_{helix-end} + \Delta G'_j|_{helix-end}$.

During suboptimal traceback, the MFE conformation for each base pair *i-j* is computed as follows in [2, 3]:

$$E_{min}(\textit{structure with i-j pair}) = V(i, j) + V(j, i+N),$$

where N is the sequence length. For RNAprob-3, we need to distinguish between two cases, based on whether *i-j* is a helix-end or stacked pair in the resulting structure, making sure that correct pseudo-energy terms are assigned. If *i-j* is a helix-end pair, we have

$$E_{min}(\textit{structure with i-j pair}) = V(i, j) + V(j, i+N) - \Delta G'_i|_{helix-end} - \Delta G'_j|_{helix-end}.$$

This formula derives from the fact that the term $\Delta G'_i|_{helix-end} + \Delta G'_j|_{helix-end}$ is counted twice, specifically, in the calculations of $V(i, j)$ and of $V(j, i+N)$, whereas it should be added only once. For the case where *i-j* is a stacked pair, we have

$$E_{min}(\textit{structure with i-j pair}) = V(i, j)+V(j, i+N)-2\times(\Delta G'_i|_{helix-end}+\Delta G'_j|_{helix-end})+\Delta G'_i|_{stacked}+\Delta G'_j|_{stacked}.$$

Performance measures

Let TP (true positives) be the number of correctly predicted base pairs, FP (false positive) the number of base pairs in the predicted structure but not in the reference structure, TN (true negative) the number of base pairs that do not exist in both predicted and reference structures and FN (false negatives) the number of base pairs in the reference structure but not in the predicted structure. Then by [4], we have

$$sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$PPV = \frac{TP}{TP + FP} \tag{2}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{3}$$

To calculate SLW-average sensitivity, PPV and MCC, we compute the sum of TPs, FPs, TNs and FNs across all RNAs. These values are then plugged into the above equations.

Simulation from a kernel density estimation

The distribution for each structural context was modeled using Gaussian kernel density estimation, as implemented in the R function *density*. The bandwidth (h) for the Gaussian kernel was selected based on the method developed by Sheather & Jones [5] (setting bw="JS" in *density*). Resulting bandwidths amounted to h=0.0367 (unpaired), 0.0159 (helix-end) and 0.0048 (stacked). For each base i, a simulated reactivity, α_i, can be generated as follows:

- Randomly sample a value, x, with replacement from the distribution corresponding to its structural context, π_i (based on the reference structure).

- Sample α_i from a normal distribution, $\mathcal{N}(x, h_{\pi_i})$.

KDE decoder

Reactivity (α) distributions were modeled using a Gaussian KDE as described above and with identical bandwidths. This resulted in the following probability sets for each structural context (π):

$$\{P(\alpha|\pi = \text{Helix-end}) \mid -2.50 \leq \alpha \leq 4.86, \alpha_n = \alpha_{n-1} + 0.0144\}$$

$$\{P(\alpha|\pi = \text{Stacked}) \mid -2.32 \leq \alpha \leq 6.88, \alpha_n = \alpha_{n-1} + 0.0180\}$$

$$\{P(\alpha|\pi = \text{Unpaired}) \mid -1.30 \leq \alpha \leq 17.12, \alpha_n = \alpha_{n-1} + 0.0361\}$$

Data transformation

SHAPE data may contain negative and zero reactivities that preclude direct log-transformation. Here, we describe a routine to generate strictly positive SHAPE profiles suitable for such transformation. This routine is implemented in R and uses the package *PearsonDS*.

1. Find $X = \{\alpha_i | \alpha_i < 0\}$ (i.e., the set of negative reactivities) across all RNAs and log-transform the absolute values of X.

   Note: The log-transformation is optional but produces a distribution with less extreme moments, therefore allowing to fit a more common Pearson distribution.

2. Determine the moments of the distribution obtained in Step 1 (using *empMoments*) .

   In our study, [μ, σ^2, skewness, kurtosis] = [-2.77, 1.54, -0.60, 5.18], corresponding to a Pearson type IV distribution.

3. For each RNA, randomly sample (using *rpearson*) n data points from a Pearson density distribution with moments computed in Step 2, where n is the total number of negatives

and zeros. Take the exponential of these n values (if a log-transform was applied in Step 1) and finally replace negative and zero reactivities with these new values.

Computing posterior probabilities of structural contexts

Probabilities for each structural context conditioned on SHAPE reactivities were determined using a Bayesian approach, similar to [6], but extended to accommodate three possible contexts. The structural context probability of a base given its observed reactivity was computed as:

$$P(\pi_i|\alpha_i) = \frac{P(\alpha_i|\pi_i) \cdot P(\pi_i)}{\sum_j P(\alpha_i|\pi_j) \cdot P(\pi_j)}.$$

The calculation requires:

- $P(\pi_i)$: The prior probability of context π_i.
- $P(\alpha_i|\pi_i)$: Reactivity likelihoods given the structure context π_i. Note that here we use a Gaussian kernel density estimate to smooth reactivity distributions. Reactivity probability densities ($P(\alpha_i|\pi_i)$) were determined using the function *kde* from the R package *ks* with a bandwidth set to 0.2.

Mock probe simulations

Let α_i be the reactivity to be simulated for base i and π_i be the corresponding structural context (in the reference structure).

Scenario 1: Mock probe reacts identically with helix-end and unpaired bases. More formally, helix-end reactivities are sampled from the unpaired SHAPE distribution, such that:

$$P(\alpha_i|\pi_i = \text{helix-end}) = P(\alpha_i|\pi_i = \text{unpaired}) \sim \exp(\alpha_i; \lambda = 1.468)$$

$$P(\alpha_i|\pi_i = \text{stacked}) \sim \text{GEV}(\alpha_i; \mu = 0.040, \sigma = 0.049, \xi = 0.763),$$

where GEV stands for the generalized extreme value distribution and exp is exponential distribution.

Scenario 2: Here, reactivities are sampled from normal distributions. G/C bases are sampled from distributions with higher variances compared to A/U bases. In more detail, if x_i is the ith base identity (i.e., $x_i \in \{A, C, G, U\}$), then reactivity α_i is simulated as follows:

$$P(\alpha_i|\pi_i = \text{stacked}, x_i) \sim \begin{cases} \mathcal{N}(\mu = 0, \sigma = 0.1), & \text{if } x_i \in \{\text{A,U}\} \\ \mathcal{N}(\mu = 0, \sigma = 1), & \text{if } x_i \in \{\text{G,C}\} \end{cases}$$

$$P(\alpha_i|\pi_i = \text{helix-end}, x_i) \sim \begin{cases} \mathcal{N}(\mu = 0.25, \sigma = 0.1), & \text{if } x_i \in \{\text{A,U}\} \\ \mathcal{N}(\mu = 0.25, \sigma = 1), & \text{if } x_i \in \{\text{G,C}\} \end{cases}$$

$$P(\alpha_i|\pi_i = \text{unpaired}, x_i) \sim \begin{cases} \mathcal{N}(\mu = 0.5, \sigma = 0.1), & \text{if } x_i \in \{\text{A,U}\} \\ \mathcal{N}(\mu = 0.5, \sigma = 1), & \text{if } x_i \in \{\text{G,C}\} \end{cases}$$

Finally, negative reactivities are set to zero. RNAprob can model negative reactivities but in RNAlin, all negatives are set to 0 before computing pseudo-energies. By setting negatives to zero, we ensure that RNAprob and RNAlin are compared on identical input information, thus allowing for a fair comparison.

For each of the two scenarios, we generated 10 replicates. For each replicate, we optimized the parameters for RNAlin and re-generated the distribution for RNAprob on a training set of RNAs, and then compared their performances on the test set (Table S4). Note that the partition into training and test sets followed Hajdin et al. (2013) to mimic the way RNAlin was recently optimized. For RNAlin, the set of m and b with the highest SLW-MCC value on the training set was chosen for each replicate. Here, for the sake of illustration, we used RNAprob-3 as a representative of the RNAprob approach.

For each scenario, a paired t-test was used to compare performances between RNAlin and RNAProb-

3. Each test compared 10 pairs of SLW-average MCCs, a pair per each replicate. For both scenarios, RNAprob consistently outperformed RNAlin (Table S5). We also compared performances on individual RNAs (70 pairs from 10 replicates for 7 test RNAs) and found that 1) for scenario 1, RNAprob outperformed RNAlin in 35 out of 70 pairs and vice versa in 20 pairs, and they performed comparably in 15 pairs; 2) for scenario 2, RNAprob outperformed RNAlin in 52 out of 70 pairs and vice versa in 6 pairs, and they performed comparably in 12 pairs. Notably, in the latter case, the SLW-average MCC score, when averaged across replicates, revealed a 15% difference in favor of RNAprob (81.06% versus 66%, Table 1), yet 6 pairs (all for RNAs shorter than 200 nt) showed higher performances with RNAlin compared to RNAprob. This volatility of performance scores at the invidual RNA level (especially for small RNAs) reinforces the need to compare schemes by SLW-averaged performances over a suffficently large test set for statistical robustness.

Statistical tests used in this study

Statistical analyses discussed in the present study were performed using several variants of conventional Student's t-tests, as implemented in the R function *t.test*.

Paired t-test: Scheme X vs. Scheme Y on real data. For both scheme X and Y, we have 23 data points, corresponding to MCC scores for the 23 RNAs in our dataset. The pairing of the t-test ensures that for an RNA, the MCC score obtained with scheme X is matched to the one obtained with scheme Y.

One-sample t-test: Performance on real data vs. simulations for scheme X. For real data, we have a single data point representing the SLW-average MCC, while for simulated data, we have a score for each of n simulations. Notably, the central limit theorem states that a normally distributed population is not a requirement given our sample size and assuming independence. Nevertheless, we confirmed the validity of the above statistical test by bootstrapping SLW-MCC scores for simulations and determining where the single data point for real data fell in the resulting distribution (data not shown).

Two-sample Welch's t-test: Scheme X vs. Scheme Y on simulated data. For both scheme X and Y, we have a set of n data points, corresponding to SLW-average MCC scores over n simulations. In this case, we assume unequal variance between the two sets.

In cross-validation studies and for the robustness-to-noise analysis, p-values were computed using only one tail of the t-distribution (one-tailed test). In this case, the test assumes directionality of the effect. For the leave-one-in analysis, a performance increase compared to the no-SHAPE control was tested. For the leave-one-out analysis and the robustness-to-noise analysis, a performance decrease compared to the performance with the entire SHAPE profile was tested. All p-values for other t-tests were computed using both tails of the t-distribution (two-tailed test).

When multiple tests were performed for the same purpose, p-values were adjusted for multiple testing using the Benjamini-Hochberg method to correct for inflated Type I errors [7]. This method is implemented in the R function $p.adjust$.

Supplementary Figures

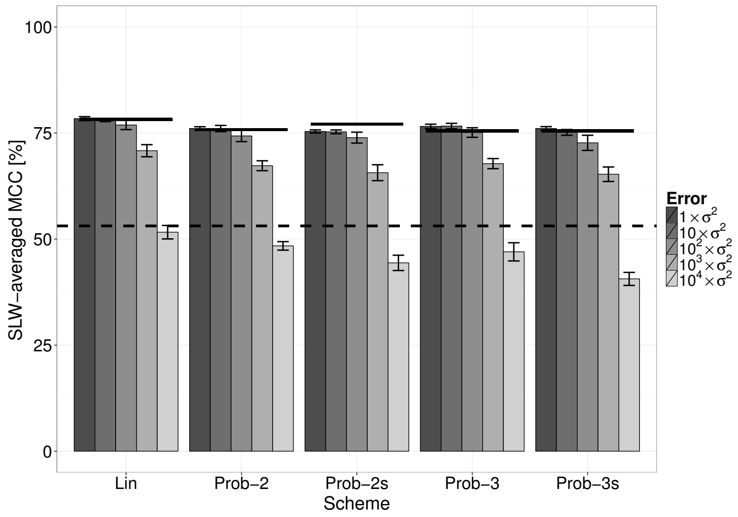

Figure S4: Performances of real data with added simulated noise. Five noise levels were applied to the data, with σ^2 denoting the variance previously observed in another dataset. Bars represent SLW-average MCCs and errors bars represent the standard deviation across simulations. The bottom dashed line indicates the performance of no-SHAPE control while scheme-specific upper lines represent performances with entire SHAPE profiles.

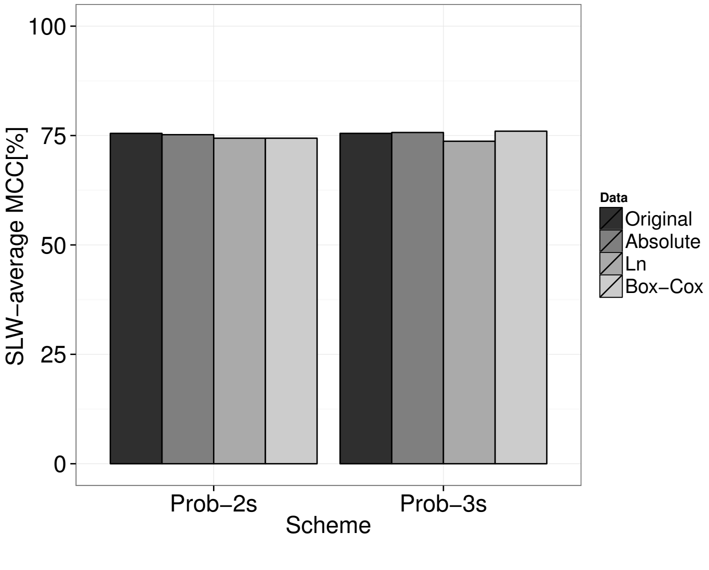

Figure S5: Performances on real data with transformation. *Original* indicates original SHAPE profiles; *Absolute* represents strictly positive profiles as obtained using random reassignment; *Ln* and *Box-Cox* represent log- and Box-Cox-transformed data using strictly positive profiles, respectively.

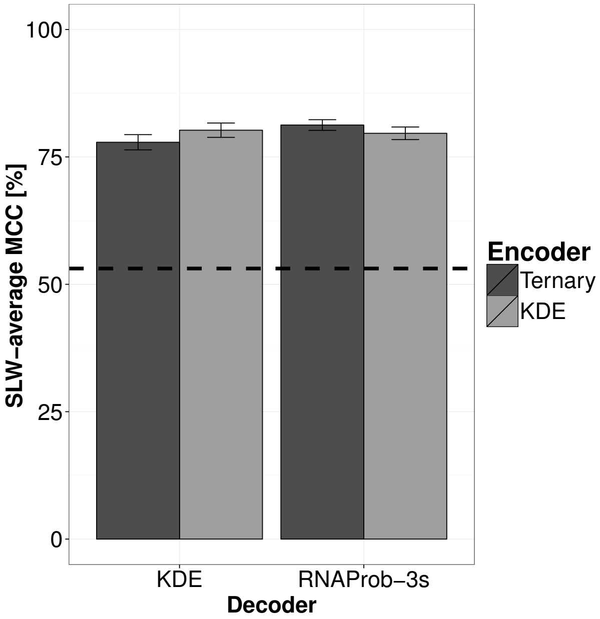

Figure S6: Performances with matching encoders-decoders. Ternary and KDE encoders indicates simulations based on the ternary model and KDE model, respectively. Decoder denotes the scheme used (See section *KDE decoder* for details on the KDE scheme). Bars represent SLW-average MCCs and errors bars represent the standard deviation across simulations (N=100). The dashed line indicates the performance of no-SHAPE control.

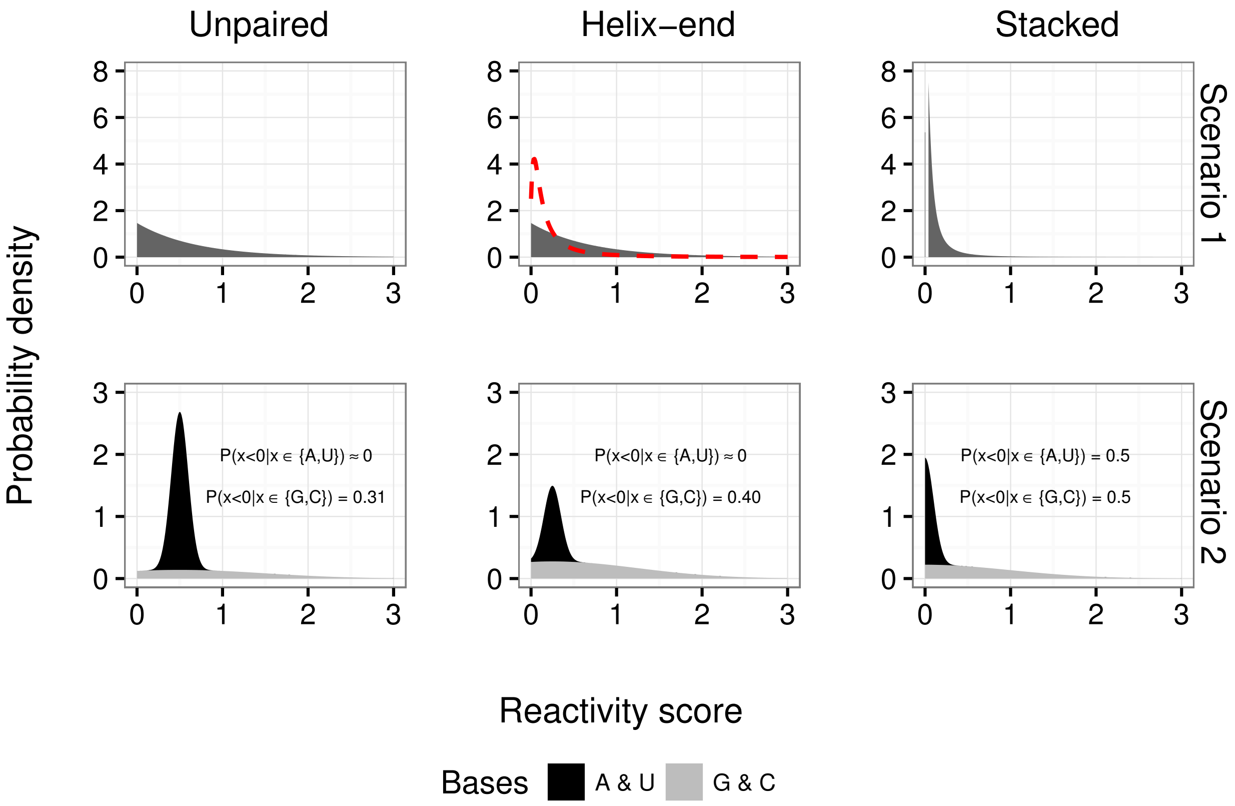

Figure S7: Reactivity distributions derived from mock probes for unpaired, helix-end and stacked bases. (Top) Scenario 1: In dark grey, the probability densities used to simulate data. The dashed red line corresponds to the real SHAPE helix-end distribution. (Bottom) Scenario 2: Distributions resulting from simulated data. Black and light grey indicate the distributions for A/U and G/C bases, respectively. Distributions were scaled such that they correspond to the average composition of the tested RNAs (i.e. 63.8%, 30.6% and 43.2% A/U bases in unpaired, helix-end and paired categories, respectively.)

Supplementary Tables

Table S1: Summary of RNA sequences used in this study.

RNA	Length	Reference of SHAPE profile
Pre-Q1 riboswitch, *B. subtilis*	34	[8]
Fluoride riboswitch, *P. syringae*	66	[8]
Adenine riboswitch, *V. vulnificus*	71	[8]
tRNA(asp), *yeast*	75	[9]
tRNA(phe), *E. coli*	76	[8]
TPP riboswitch, *E. coli*	79	[8]
SARS corona virus pseudoknot	82	[8]
cyclic-di-GMP riboswitch, *V. cholerae*	97	[8]
SAM I riboswitch, *T. tengcongensis*	118	[8]
5S rRNA, *E. coli*	120	[8]
M-Box riboswitch, *B. subtilis*	154	[8]
P546 domain, bI3 group I intron	155	[9]
Lysine riboswitch, *T. maritima*	174	[8]
Group I intron, *Azoarcus sp.*	214	[8]
Hepatitis C virus IRES domain	336	[8]
Group II intron, *O. iheyensis*	412	[8]
Group I Intron, *T. thermophila*	425	[8]
$5'$ domain of 23S rRNA, *E. coli*	511	[8]
$5'$domain of 16S rRNA, *E. coli*	530	[8]
16S rRNA, *H. volcanii*	1474	[10]
16S rRNA, *C. difficile*	1503	[10]
16S rRNA, *E. coli*	1542	[9]
23S rRNA, *E. coli*	2904	[9]

Table S2: Performances on real data

		noSHAPE			RNAlin			RNAprob-2			RNAprob-3		
RNA	Length	Sen.	PPV	MCC	Sen.	PPV	MCC	Sen.	PPV	MCC	Sen.	PPV	MCC
Pre-Q1 riboswitch, *B. subtilis*	34	62.5	100	78.8	62.5	100	78.8	62.5	83.3	71.8	62.5	83.3	71.8
Fluoride riboswitch, *P. syringae*	66	56.2	64.3	59.9	62.5	71.4	66.6	56.2	64.3	59.9	56.2	64.3	59.9
Adenine riboswitch, *V. vulnificus*	71	100	100	100	100	100	100	100	100	100	100	100	100
tRNA(asp), *yeast*	75	76.2	76.2	76.0	76.2	76.2	76.0	100	100	100	100	100	100
tRNA(phe), *E. coli*	76	95.2	100	97.6	100	100	100	100	100	100	100	100	100
TPP riboswitch, *E. coli*	79	77.3	85.0	80.9	95.5	87.5	91.3	95.5	87.5	91.3	95.5	87.5	91.3
SARS corona virus pseudoknot	82	69.2	90.0	78.8	69.2	75.0	71.8	69.2	75.0	71.8	69.2	75.0	71.8
cyclic-di-GMP riboswitch, *V. cholerae*	97	75.0	77.8	76.2	89.3	86.2	87.7	75.0	72.4	73.5	75.0	72.4	73.5
SAM I riboswitch, *T. tengcongensis*	118	74.4	80.6	77.3	76.9	85.7	81.1	53.9	72.4	62.3	64.1	78.1	70.6
5S rRNA, *E. coli*	120	28.6	25.0	26.3	85.7	76.9	81.1	97.1	91.9	94.5	97.1	91.9	94.5
M-Box riboswitch, *B. subtilis*	154	87.5	91.3	89.3	87.5	91.3	89.3	54.2	59.1	56.4	54.2	61.9	57.8
P546 domain, bI3 group I intron	155	42.9	44.4	43.4	94.6	96.4	95.5	76.8	87.8	82.0	76.8	87.8	82.0
Lysine riboswitch, *T. maritima*	174	68.2	72.9	70.4	69.8	78.6	74.0	69.8	75.9	72.7	69.8	75.9	72.7
Group I intron, *Azoarcus sp.*	214	66.7	68.8	67.7	77.8	81.7	79.6	76.2	82.8	79.3	69.8	75.9	72.7
Hepatitis C virus IRES domain	336	39.4	38.0	38.6	79.8	86.5	83.0	78.8	86.3	82.5	79.8	86.5	83.0
Group II intron, *O. iheyensis*	412	88.6	97.5	93.0	74.2	84.5	79.2	76.5	89.4	82.7	84.1	93.3	88.5
Group I Intron, *T. thermophila*	425	77.9	68.9	73.2	86.3	81.9	84.0	81.7	84.2	82.9	84.0	84.6	84.3
$5'$ domain of 23S rRNA, *E. coli*	511	87.4	69.8	78.1	89.9	74.3	81.7	89.1	76.3	82.4	85.7	72.3	78.7
$5'$domain of 16S rRNA, *E. coli*	530	56.8	52.5	54.5	89.2	80.0	84.5	88.5	84.5	86.5	91.2	84.4	87.7
16S rRNA, *H. volcanii*	1474	56.5	52.1	54.3	77.5	75.7	76.6	76.9	76.5	76.7	76.0	75.0	75.5
16S rRNA, *C. difficile*	1503	39.0	37.0	38.0	74.8	74.0	74.4	68.5	72.2	70.3	69.8	70.6	70.2
16S rRNA, *E. coli*	1542	36.8	33.8	35.2	79.7	75.4	77.5	68.7	68.3	68.5	78.1	73.6	75.8
23S rRNA, *E. coli*	2904	53.1	48.1	50.5	76.6	73.6	75.1	76.1	74.6	75.4	77.3	74.9	76.1
Average		65.9	68.4	66.9	81.5	83.2	82.1	77.9	81.1	79.3	79.0	81.3	79.9
SLW-average		54.8	51.4	53.1	79.2	77.2	78.2	75.4	76.3	75.8	77.5	76.6	77.1

Table S2: Performances on real data (Cont.)

		RNAprob-2s			RNAprob-3s		
RNA	Length	Sensitivity	PPV	MCC	Sensitivity	PPV	MCC
Pre-Q1 riboswitch, *B. subtilis*	34	62.5	83.3	71.8	37.5	50.0	42.6
Fluoride riboswitch, *P. syringae*	66	56.2	64.3	59.9	56.2	69.2	62.1
Adenine riboswitch, *V. vulnificus*	71	100	100	100	100	100	100
tRNA(asp), *yeast*	75	57.1	63.2	59.8	52.4	52.4	52.0
tRNA(phe), *E. coli*	76	76.2	76.2	76.0	100	100	100
TPP riboswitch, *E. coli*	79	95.5	87.5	91.3	95.5	87.5	91.3
SARS corona virus pseudoknot	82	69.2	75.0	71.8	69.2	90.0	78.8
cyclic-di-GMP riboswitch, *V. cholerae*	97	75.0	72.4	73.5	96.4	93.1	94.7
SAM I riboswitch, *T. tengcongensis*	118	64.1	78.1	70.6	64.1	78.1	70.6
5S rRNA, *E. coli*	120	97.1	91.9	94.5	85.7	76.9	81.1
M-Box riboswitch, *B. subtilis*	154	68.8	68.8	68.6	68.8	67.3	67.9
P546 domain, bI3 group I intron	155	69.6	86.7	77.6	76.8	87.8	82.0
Lysine riboswitch, *T. maritima*	174	69.8	78.6	74.0	68.2	75.4	71.6
Group I intron, *Azoarcus sp.*	214	63.5	81.6	71.9	73.0	88.5	80.3
Hepatitis C virus IRES domain	336	78.8	86.3	82.5	77.9	85.3	81.5
Group II intron, *O. iheyensis*	412	76.5	91.8	83.8	68.2	79.0	73.3
Group I Intron, *T. thermophila*	425	80.9	84.1	82.5	80.9	80.3	80.6
5′ domain of 23S rRNA, *E. coli*	511	90.8	77.1	83.7	75.6	64.8	70.0
5′domain of 16S rRNA, *E. coli*	530	88.5	84.5	86.5	88.5	84.5	86.5
16S rRNA, *H. volcanii*	1474	75.5	78.1	76.8	76.0	75.8	75.9
16S rRNA, *C. difficile*	1503	69.6	73.5	71.5	66.8	70.0	68.4
16S rRNA, *E. coli*	1542	66.7	66.6	66.7	76.2	74.2	75.2
23S rRNA, *E. coli*	2904	74.9	75.8	75.4	75.7	76.1	75.9
Average		75.1	79.4	77.0	75.2	78.5	76.6
SLW-average		74.4	76.7	75.5	75.0	76.1	75.5

Table S3: Quintile ranges for SHAPE reactivities

	Quintile					
RNA	0%	20%	40%	60%	80%	100%
Pre-Q1 riboswitch, *B. subtilis*	-0.429	-0.130	-0.013	0.048	0.266	5.206
Fluoride riboswitch, *P. syringae*	-0.354	-0.025	0.034	0.080	0.673	1.951
Adenine riboswitch, *V. vulnificus*	0	0.050	0.100	0.150	0.375	5.175
tRNA(asp), *yeast*	0.035	0.085	0.131	0.225	0.409	1.830
tRNA(phe), *E. coli*	0	0	0.010	0.020	0.137	1.607
TPP riboswitch, *E. coli*	0	0.032	0.053	0.075	0.177	1.548
SARS corona virus pseudoknot	-2.456	-0.355	-0.046	0.210	0.602	6.426
cyclic-di-GMP riboswitch, *V. cholerae*	0	0	0.018	0.028	0.210	2.065
SAM I riboswitch, *T. tengcongensis*	-0.041	0.047	0.186	0.419	0.869	3.864
5S rRNA, *E. coli*	-0.083	0.050	0.134	0.381	0.875	3.255
M-Box riboswitch, *B. subtilis*	0	0.081	0.162	0.357	0.775	17.01
P546 domain, bI3 group I intron	0	0.024	0.135	0.367	0.798	3.713
Lysine riboswitch, *T. maritima*	0	0.078	0.195	0.351	0.857	9.045
Group I intron, *Azoarcus sp.*	-0.145	0.070	0.163	0.393	0.847	7.471
Hepatitis C virus IRES domain	-0.392	0.094	0.198	0.384	0.778	4.971
Group II intron, *O. iheyensis*	-0.581	-0.026	0.145	0.363	0.810	3.281
Group I Intron, *T. thermophila*	-0.934	-0.032	0.085	0.238	0.719	5.052
$5'$ domain of 23S rRNA, *E. coli*	0	0.027	0.097	0.263	0.783	3.539
$5'$domain of 16S rRNA, *E. coli*	0	0.028	0.107	0.294	0.775	5.047
16S rRNA, *H. volcanii*	-1.449	-0.005	0.083	0.269	0.725	6.758
16S rRNA, *C. difficile*	-1.032	0.003	0.135	0.314	0.750	4.311
16S rRNA, *E. coli*	0	0.030	0.104	0.278	0.633	5.000
23S rRNA, *E. coli*	0	0.048	0.120	0.295	0.703	7.419
All RNAs combined	-2.456	0.028	0.111	0.290	0.720	17.01

Table S4: RNA sequences in the training set and test set used in the mock probe analyses.

Training set	Test set
Pre-Q1 riboswitch, *B. subtilis*	Fluoride riboswitch, *P. syringae*
tRNA(asp), *yeast*	Adenine riboswitch, *V. vulnificus*
TPP riboswitch, *E. coli*	tRNA(phe), *E. coli*
SARS corona virus pseudoknot	5S rRNA, *E. coli*
cyclic-di-GMP riboswitch, *V. cholerae*	16S rRNA, *H. volcanii*
SAM I riboswitch, *T. tengcongensis*	16S rRNA, *E. coli*
M-Box riboswitch, *B. subtilis*	23S rRNA, *E. coli*
P546 domain, bI3 group I intron	
Lysine riboswitch, *T. maritima*	
Group I intron, *Azoarcus sp.*	
Hepatitis C virus IRES domain	
Group II intron, *O. iheyensis*	
Group I Intron, *T. thermophila*	
5′ domain of 23S rRNA, *E. coli*	
5′domain of 16S rRNA, *E. coli*	
16S rRNA, *C. difficile*	

Table S5: The SLW-MCC value (%) on the test set for each of the two scenarios.

	Scenario 1		Scenario 2	
Replicate	RNAprob-3	RNAlin	RNAprob-3	RNAlin
1	78.2	78	83.2	66.6
2	80.3	72.5	82.9	66.5
3	81	76.7	80	63.8
4	80.7	76.2	81.1	70.7
5	78.5	71.1	80.6	67.3
6	78.4	72.8	79.4	67.2
7	81	75.3	80.9	63.2
8	78.7	74.4	80.4	60.8
9	79.8	72.6	81.9	66.4
10	81.4	76.1	80.2	67.5

References

[1] Reuter JS, Mathews DH. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinf* **11:** 129.

[2] Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci* **101:** 7287–7292.

[3] Zuker M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* **244:** 48–52.

[4] Gardner PP, Giegerich R. 2004. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinf* **5** 140.

[5] Sheather SJ, Jones MC. 1991. A reliable data-based bandwidth selection method for kernel density estimation. *J R Stat Soc Series B* **53:** 683–690.

[6] Bindewald E, Wendeler M, Legiewicz M, Bona MK, Wang Y, Pritt MJ, Le Grice SFJ, Shapiro BA. 2011. Correlating SHAPE signatures with three-dimensional RNA structures. *RNA* **17** 1688–1696.

[7] Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B*, **57:** 289–300.

[8] Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. 2013. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc Natl Acad Sci* **110:** 5498–5503.

[9] Deigan KE, Li TW, Mathews DH, Weeks KM. 2009. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci* **106:** 97–102.

[10] Lavender CA, Lorenz R, Zhang G, Tamayo R, Hofacker IL, Weeks KM. 2015. Model-free RNA sequence and structure alignment informed by SHAPE probing reveals a conserved alternate secondary structure for 16S rRNA. *PLoS Comput Biol* **11:** e1004126.

[11] Eddy SR. 2014. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annu Rev Biophys* **43:** 433–456.